

# Least Absolute Gradient Selector: variable selection via Pseudo-Hard Thresholding

Kun Yang\*

March 2, 2013

## Abstract

In this paper, we propose a new approach, called the **LAGS**, short for “least absolute gradient selector”, to this challenging yet interesting problem by mimicking the discrete selection process of  $l_0$  regularization in linear regression. To estimate  $\beta$  under the influence of noise, we consider, nevertheless, the following convex program

$$\hat{\beta} = \arg \min \frac{1}{n} \|X^T(y - X\beta)\|_1 + \lambda_n \sum_{i=1}^p w_i(y; X; n) |\beta_i|$$

$\lambda_n > 0$  controls the sparsity and  $w_i > 0$  dependent on  $y, X$  and  $n$  is the weights on different  $\beta_i$ ;  $n$  is the sample size. Surprisingly, we shall show in the paper, both geometrically and analytically, that LAGS enjoys two attractive properties: (1) LAGS demonstrates discrete selection behavior and hard thresholding property as  $l_0$  regularization by strategically chosen  $w_i$ , we call this property “*pseudo-hard thresholding*”; (2) Asymptotically, LAGS is consistent and capable of discovering the true model; nonasymptotically, LAGS is capable of identifying the sparsity in the model and the prediction error of the coefficients is bounded at the noise level up to a logarithmic factor— $\log p$ , where  $p$  is the number of predictors.

Computationally, LAGS can be solved efficiently by convex program routines for its convexity or by simplex algorithm after recasting

---

\*Kun Yang (Email: kunyang@stanford.edu) is a PhD student at Institute for Computational and Mathematical Engineering, Stanford University. Kun Yang is partially supported by General Wang Yaowu Fellowship.

it into a linear program. The numeric simulation shows that LAGS is superior compared to soft-thresholding methods in terms of mean squared error and parsimony of the model.

## 1 Introduction

One of the most widely used model in statistics is the linear regression. In many applications, scientists are interested in estimating a mean response  $X\beta$  from the data  $y = (y_1, y_2, \dots, y_n)$ . The  $p$ -dimensional parameter of interest  $\beta$  are estimated from the linear model

$$y = \alpha + X\beta + \epsilon \quad (1)$$

where  $\alpha$  is the intercept,  $\epsilon$  the noise. A common assumption is that  $\epsilon$  is Gaussian with  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ , but this is not an essential requirement as our methods are applicable to other types of noise which have heavier tails.

Scientists usually have no information of the underlying models, a large number of predictors are chosen in initial stage to attenuate the possible bias and variance as well as to enhance predictability. Regression procedures capable of identifying the explanatory variables (others are set to be or shrunk to 0) and obtain good estimates of the response play a pivotal role. Ideally, Best Subset regression such AIC [1],  $C_p$  [12], BIC [14] and RIC [10] achieve the trade-off between model complexity and goodness of fit. These estimators are essentially least square penalized by  $l_0$  norm of  $\beta$  with different control coefficients. A standard formulation of  $l_0$  penalized least square is

$$(\hat{\alpha}, \hat{\beta}) = \arg \min \frac{1}{2} \|y - \alpha \mathbf{1} - X\beta\|_2^2 + \lambda \|\beta\|_0 \quad (2)$$

$\lambda$  is the control coefficient. For the rest of the paper, we assume that columns of  $X$  are standardized and  $y$  is centered. Under this setting  $\alpha = 0$ , hence we omit it.

$l_0$  penalized least square has the hard thresholding property that keeps the large coefficient intact while sets small ones to be zero. However, unfortunately, solving (2) is a discrete process which needs to enumerate all the possible subset of  $\beta$ ; the combinatorial nature of  $l_0$  norm limits the application of (2) when the number of predictors is large. As a compromise, convex relaxation to  $l_1$  norm such as the Lasso [16] is a widely used technique for

simultaneously estimation and variable selection. The Lasso is

$$\hat{\beta} = \arg \min \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad (3)$$

The  $l_1$  penalty shrinks  $\beta$  towards zero, and also sets many coefficients to be exactly zero. Thus, the Lasso often regards as the substitute of (2).

The Lasso and the subsequently appeared methods, such as the LARS [5], elastic-net [20], adaptive Lasso [19], Dantzig Selector [2] to name a few, closely relate to the soft thresholding in signal processing [4] in contrast to hard thresholding as in (2). Since the soft thresholding both shrinks and selects, it often results in a model more complicated than the true model in its effort to spread the penalty among the predictors. As a greedier attempt, SCAD [7, 9, 8] and SparseNet [13] penalize the loss function by non-convex penalties. Similar to (3),  $\beta$  is estimated by

$$\hat{\beta} = \arg \min \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \sum_{i=1}^p P(|\beta_i|; \lambda; \gamma) \quad (4)$$

where  $P(|\beta_i|; \lambda; \gamma)$  defines a family of penalty functions concave in  $\beta$ , and  $\lambda$  and  $\gamma$  controls the sparsity and concavity. It has been shown that (4) enjoys better variable selection properties compared to  $l_1$  relaxation; whereas, both algorithms cannot assure to find the global optimal.

In this paper, we propose a new approach, called the **LAGS**, short for “least absolute gradient selector”, to this challenging yet interesting problem by mimicking the discrete selection process of  $l_0$  regularization. To estimate  $\beta$  under the influence of noise, we consider, nevertheless, the following convex program

$$\hat{\beta} = \arg \min \frac{1}{n} \|X^T(y - X\beta)\|_1 + \lambda_n \sum_{i=1}^p w_i(y; X; n) |\beta_i| \quad (5)$$

$\lambda_n > 0$  controls the sparsity and  $w_i > 0$  dependent on  $y, X$  and  $n$  is the weights on different  $\beta_i$ ;  $n$  is the sample size. Surprisingly, we shall show in the following sections, both geometrically and analytically, that (5) demonstrates discrete selection behavior as  $l_0$  penalty and hard thresholding property by strategically chosen  $w_i$ , we call this property “*pseudo-hard thresholding*”. The graphical comparison of hard thresholding and pseudo-hard thresholding is given in prostate cancer example in section 7.

The rest of the paper is organized as follows: section 2 presents the motivation of LAGS and connects the ideas with previous work; section 3 establishes the theorem regarding the properties of LAGS and highlights the “pseudo-hard thresholding”; section 4 shows the potential problems associated with the Dantzig Selector and provides a neat way to choose  $w_i$ ; section 5 are the proofs of the theorems; section 6 discusses the computational issue of how to solve LAGS; section 7 demonstrates it by numeric examples; discussion and future work are in section 8.

## 2 The LAGS

### 2.1 Insight from orthonormal case

The properties of hard thresholding can be better understood when the design  $X$  is orthonormal, i.e.,  $X^T X = I$ . The solution for (2) is

$$\hat{\beta}_j^{l_0} = \beta_j^o \mathbf{I}(|\beta_j^o| \geq \lambda) \quad (6)$$

as a reference, the Lasso solution is

$$\hat{\beta}_j^{lasso} = \text{sign}(\beta_j^o)(|\beta_j^o| - \lambda)_+ \quad (7)$$

where  $\beta^o = X^T y$  is the ordinary least square (OLS) estimate. Notice that the hard thresholding operator (6) is discontinuous with the jump at  $|\beta_j^o| = \lambda$ , while the soft thresholding operator (7) is continuous. If we assume  $|\beta_1^o| \geq \dots |\beta_{p_0}^o| > |\beta_{p_0+1}^o| \dots \geq |\beta_p^o|$ , another property of (6) is that any  $\lambda \in (|\beta_{p_0}^o|, |\beta_{p_0+1}^o|)$  will keep the first  $p_0$  coefficients. Based on this property, we define “pseudo-hard thresholding” as

**Definition 2.1.** A penalized regression has “pseudo-hard thresholding” property if there exist some intervals of  $R^+$ , such that changing sparsity parameter  $\lambda$  in these intervals will keep the coefficients  $\beta$  unaltered.

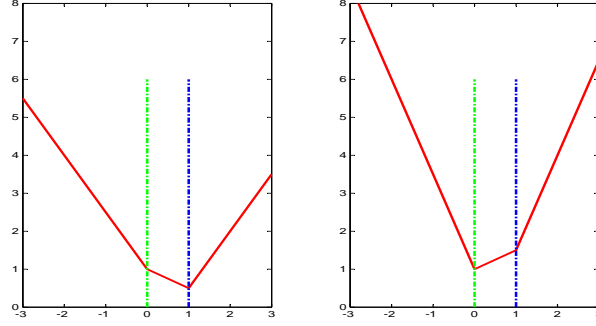
In order to achieve pseudo-hard thresholding in convex world, consider function of  $\theta$

$$f(\theta; z) = |z - \theta| + \frac{\gamma}{\lambda} |\theta| \quad (8)$$

Minimizing  $f(\theta; z)$ , the solution is

$$\hat{\theta} = \begin{cases} z & \gamma \leq \lambda \\ 0 & \gamma > \lambda \end{cases} \quad (9)$$

Figure 1: Plot of  $f(\theta; z)$ . The left panel has  $\gamma/\lambda < 1$ ; the right panel  $\gamma/\lambda > 1$ . It is obvious that the minimum is obtained at  $x = 1$  on the left, while  $x = 0$  on the right



$\hat{\theta}$  is a discontinuous function of  $\gamma$ , with the breaking point at  $\gamma = \lambda$ . Motivated by this special property of (8), we formulate LAGS as in (5). As a matter of fact, the idea to minimize  $\|X^T(y - X\beta)\|$  is pioneered by [2] in Dantzig Selector

$$\hat{\beta} = \arg \min \|\beta\|_1 \quad \text{subject to} \quad \|X^T(y - X\beta)\|_\infty \leq t \quad (10)$$

which can be written equivalently as

$$\hat{\beta} = \arg \min \|X^T(y - X\beta)\|_\infty \quad \text{subject to} \quad \|\beta\|_1 \leq t \quad (11)$$

One reason why  $\|X^T(y - X\beta)\|$  should be small given in [2] is that a good estimate of  $\beta$  should be independent of orthogonal transformations. Another more compelling yet insightful argument is that in OLS one needs to minimize  $f_{OLS} = \frac{1}{2}\|y - X\beta\|_2^2$ , its gradient is  $\nabla f_{OLS} = -X^T(y - X\beta)$ ; solving  $\nabla f_{OLS} = 0$  results in OLS estimates, hence, one can expect that a good  $\hat{\beta}$  should shrink  $\nabla f_{OLS}$  towards  $\mathbf{0}$ . In order to incorporate (8),  $l_1$  is chosen here—from where the name “LAGS” comes; we will show in the following sections that this choice of norm can set some elements of the gradient to zero, which means unrestricted coefficients for them, like in  $l_0$  penalty. Moreover, if we instead consider the absolute deviance by substituting  $\|X^T(y - X\beta)\|_1$  with  $\|y - X\beta\|_1$ , it is the LAD-Lasso [17]; but LAD-Lasso has no hard thresholding property even in orthonormal design case.

## 2.2 The weight $w_i$ matters

In the orthonormal case, (5) becomes

$$\hat{\beta}_i = \arg \min |\beta_i^o - \hat{\beta}_i| + \lambda w_i |\hat{\beta}_i|, i = 1, \dots, p \quad (12)$$

One heuristic to choose  $w_i$  is to consider the correlation  $c_i$  between  $y$  and  $x_i$ —the  $i$ th column of  $X$ : if  $|c_i|$  is large, which means  $i$ th predictor may be a good explanatory variable, hence  $\beta_i$  should be penalized less; otherwise, it should be penalized more. Thus, we set

$$w_i = \frac{1}{|c_i|} \quad (13)$$

with a little abuse of notation,  $w_i = \infty$  when  $c_i = 0$ . Without loss of generality, we assume that  $|c_1| \geq \dots |c_{p_0}| > \dots \geq |c_p|$ , which implies  $w_1 \leq \dots w_{p_0} < \dots \leq w_p$ . By choosing  $\lambda$ , s.t.  $w_{p_0+1}^{-1} < \lambda < w_{p_0}^{-1}$ , we have  $\hat{\beta}_i = \beta_i^o, i = 1, \dots, p_0$ ;  $\hat{\beta}_i = 0, i = p_0 + 1, \dots, p$ , which is pseudo-hard thresholding and  $\hat{\beta}$  is identical with (6).

To push this heuristic further, we notice that  $c_i$  is the coefficient of OLS in the orthonormal design case. This suggests that we can choose  $w_i^{-1}$  as absolute value of OLS coefficients  $\beta_i^o$ , see section 4 for detail.

## 3 Properties of LAGS

In section 2, some properties of LAGS are demonstrated in the simplest case. These properties are not incidental. To formally state our results, we decompose the regression coefficient as  $\beta = (\beta^{(1)}, \beta^{(2)})$ , where  $\beta^{(1)} = (\beta_1, \dots, \beta_{p_0})$  corresponds to the true parameters and  $\beta^{(2)} = (\beta_{p_0+1}, \dots, \beta_p)$  are redundant; the columns of  $X$  are decomposed alike. Furthermore, define

$$a_n = \max_{1 \leq j \leq p_0} \{w_j^n\} \quad (14)$$

and

$$b_n = \min_{p_0+1 \leq j \leq p} \{w_j^n\} \quad (15)$$

We assume three conditions:

- (a)  $y = x^{(1)}\beta^{(1)} + \epsilon$ , where  $\epsilon$  is noise with mean 0 and variance  $\sigma^2$ .

(b)  $C_n = \frac{1}{n}X^TX \rightarrow C$  and  $C_n = \begin{bmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{bmatrix}$ ,  $C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$ , where  $C_n$  and  $C$  are positive definite matrices.

(c)  $\|C_{11}^{-1}C_{12}\|_\infty \leq 1 - \eta$  and  $\|(C_{11}^n)^{-1}C_{12}^n\|_\infty \leq 1 - \eta_n$ , where  $\eta$  and  $\eta_n$  are positive constants. This condition is established in [19] and also called Irrepresentable Condition in [18].

$C$  is actually the covariance matrix of predictors. Consider set  $\Omega = \{s \in \mathbb{R}^p : \|s\|_\infty = 1, \max\{|s_1|, \dots, |s_{p_0}|\} = 1\}$ , which is closed and contains in the unit sphere under  $l_\infty$  norm, hence  $\Omega$  is a compact set. Let us define

$$\gamma = \min_{s \in \Omega} \{ \| [C_{11}, C_{12}](s^{(1)}, s^{(2)})^T \|_\infty \} \quad (16)$$

here we partition  $s$  as above. It is obvious that  $\|s^{(1)}\|_\infty = 1$  and  $\|s^{(2)}\|_\infty \leq 1$ , thus  $\|[I, C_{11}^{-1}C_{12}]s\|_\infty \geq \|s^{(1)}\|_\infty - \|C_{11}^{-1}C_{12}s^{(2)}\|_\infty \geq 1 - (1 - \eta) = \eta$ .  $\gamma > 0$  is trivial by noting that  $\|C_{11}v\|_\infty = 0 \Leftrightarrow v = 0$ .

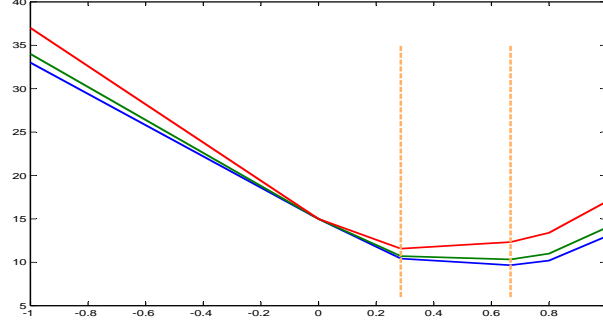
**Theorem 3.1.** *We can find  $\tilde{a}, \tilde{b}$  dependent on  $C$ , typically,  $\tilde{a} = \gamma, \tilde{b} = \|C\|_\infty$ , such that if  $\lim \lambda_n a_n < \tilde{a}$  and  $\lim \lambda_n b_n > \tilde{b}$ , then LAGS is consistent and has pseudo-hard thresholding property.*

The proof is given in Section 5. Theorem 3.1 gives another hint that the weights will play an important role for the effectiveness of (5). To clarify the pseudo-hard thresholding property, it is helpful to interpret LAGS geometrically.  $l(\beta; \lambda_n) = \frac{1}{n}\|X^T(y - X\beta)\|_1 + \lambda_n \sum_{i=1}^p w_i(y; X; n)|\beta_i|$  is piecewise linear in  $(p+1)$ -dimensional space. We assume there are no flat regions—in which  $l(\beta)$  are constant, then  $l(\beta)$  must obtain its minimum on some breaking point  $\beta'$  otherwise there exist descent directions (consider the simplex algorithm). When  $n$  is fixed, as we change  $\lambda_n$  with a tiny amount, the value in each breaking point may change, but it is possible that no values on other breaking points catch up  $l(\beta')$  after the change; if this is the case,  $\beta'$  will still be the minimizer even though the  $\lambda_n$  is different. To illustrate this point graphically, let us consider a toy 1-dimensional example  $y = (7, 2, 4, 2)^T$ ,  $X = (2, 3, 5, 7)^T$

$$l(\beta) = |7 - 2\beta| + |2 - 3\beta| + |4 - 5\beta| + |2 - 7\beta| + \lambda|\beta|$$

When  $\lambda = 1, 2, 5$ ,  $l(\beta; 1)$  and  $l(\beta; 2)$  are both minimized at  $\beta = 2/3$ , while  $l(\beta; 5)$  at  $\beta = 2/7$  as shown in Figure 2. It implies even though  $\lambda$  increases

Figure 2: Graphical illustration of pseudo-hard thresholding and discrete selection process in one dimension. The blue line is  $\lambda = 1$ ; the green line  $\lambda = 2$ ; the red line  $\lambda = 5$ . The optimal  $\beta$ s are indicated by the dotted vertical line.



from 1 to 2, the values at other breaking points do not catch up  $l(2/3)$ , but when increases to 5,  $l(2/3)$  is caught up by  $l(2/7)$ . The consistency can be understood in a similar fashion: as  $n$  increases, the number of breaking points increases exponentially, which provides more candidates to solve (5). As a consequence, the probability to discover the true model increases.

The discrete selection nature of LAGS can also be understood under this framework:  $\hat{\beta}$  only moves from one breaking point to another and the breaking points are scattered in  $p$ -dimensional space, which implies  $\hat{\beta}$  is selected discretely. Fortunately, this feat of LAGS can be achieved by solving a tractable convex program rather than enumerating all the possible breaking points as  $l_0$  regularized regression.

## 4 How to choose weights

### 4.1 Dantzig Selector

The performance of Dantzig Selector (DS) and its similarities with Lasso and LARS are discussed in [6]. In their several numerical studies, the coefficient profiles of DS seem to be wilder and have some erratic fluctuations; the prediction accuracy is also inferior. DS can be re-expressed in penalized form as

$$\hat{\beta} = \arg \min \|X^T(y - X^T\beta)\|_{\infty} + \lambda \|\beta\|_1 \quad (17)$$



We denote the penalized form as  $l_{DS}(\beta)$ , it is also a piecewise linear function. We argue that one possible reason for these properties in [6] is that the penalties on the coefficients are uniform.

*Example.* If  $l_{DS}(\beta) = \max\{|1 - \beta_1|, |2 - \beta_2|\} + \lambda(|\beta_1| + |\beta_2|)$  by carefully chosen  $y$  and  $X$ . As discussed in Section 3, the piecewise linear function will obtain its minimum at one of its breaking points. In this example, there are total four such points:  $(0, 0)$ ,  $(0, 2)$ ,  $(1, 0)$ ,  $(1, 2)$ , thus

$$\min l_{DS}(\beta) = \min\{2, 1 + 2\lambda, 2 + \lambda, 3\lambda\}$$

$$\min l_{DS} = \begin{cases} 2 & \text{if } \lambda < 2/3, \beta = (0, 0) \\ 2 & \text{if } \lambda = 2/3, \text{there are infinit } \beta s \\ 3\lambda & \text{if } \lambda > 2/3, \beta = (1, 2) \end{cases}$$

So DS will threshold both  $\beta_1$  and  $\beta_2$  or keep both intact unless  $\lambda = 2/3$ , where the solution is not unique. LAGS has the similar behavior if the weights are uniform, i.e., if we take  $\lambda w_i$ s are equal in (12). However, Theorem 3.1 requires the weights converge to different values for true and noisy predictors; hence, we conjecture that if we choose the weights for DS as with LAGS, the counterintuitive behaviors can be eschewed in some extend. We show that numerically in section 8.

## 4.2 How to choose $w_i$

Imposing different weights on the coefficients to enhance predicability is discussed in [19] and [17]. Both suggest to use the inverse of ordinary least square estimate as the weight. In the orthonormal case, the adaptive lasso estimates for  $\theta$  are obtained by

$$\hat{\beta}_j^{adaptive} = \arg \min_{\beta} \frac{1}{2}(\beta_j^o - \beta)^2 + \lambda \frac{1}{|\beta_j^o|^\gamma} |\beta|$$

where  $j = 1, 2, \dots, p$ . Therefore,  $\hat{\beta}_j^{adaptive} = \text{sign}(\beta_j^o)(|\beta_j^o| - \frac{\lambda}{|\beta_j^o|^\gamma})_+$ . Compared with Lasso solution (7), adaptive lasso shrinks  $\beta_j^o$  towards 0 less for larger  $\beta_j^o$ ; as a result, it is shown that it introduces less bias than Lasso. The weight derived in LAD-Lasso [17] is based on the Bayesian perspective: if each coefficient is double-exponentially distributed with location 0 and scale  $\lambda_i$ , then the log-likelihood of the posterior is:

$$\sum_{i=1}^n \log f(\epsilon_i) + \sum_{i=1}^p \lambda_i |\beta_i| - \log(\lambda_i) + \text{constant}$$

minimize it with respect to  $\lambda_i$ , which leads to  $\lambda_i = 1/|\beta_i|$ . However, we do not have the oracle to know  $\beta_i$  in advance; hence, at first step, we need a coarse estimate of  $\beta_i$ , a natural choice will be the  $\beta_{OLS}$ . Surprisingly, in what follows in this section, it is shown that this is also a good choice for the weights in LAGS.

**Theorem 4.1.** *If we choose  $w_i = 1/|\beta_{OLS}|_i$ , then asymptotically, conditions for  $a_n$  and  $b_n$  in Theorem 3.1 is satisfied, where  $\beta_{OLS} = (X^T X)^{-1} X^T y$ .*

What if  $p > n$  which is not addressed in [17, 19]? The number of variables  $p$  is larger than the sample size  $n$  frequently arises in applications. The ordinary least square fails because the solution of it is not unique. Nevertheless, ridge regression is a shrinkage estimator that can enhance the predictability of the model, so we use ridge solution as the weights for LAGS; moreover, the ridge regression solution can be obtained by simply solving the similar equation, namely  $\beta_{ridge} = (X^T X + \phi I)^{-1} y$ , where  $\phi$  is some positive constant; then  $w_i = 1/|\beta_{ridge}|_i$ .

There are two nice properties associated with our choice of  $w_i$ :

- (1) Without loss of generality, let us consider the scenario that there are several groups of identical predictors, the ridge estimate of  $\beta$  will always put equal weights on the identical predictors. The reason is that the optimization problem

$$\min \alpha_1^2 + \alpha_2^2 + \dots + \alpha_k^2 \quad \text{subject to} \quad \alpha_1 + \alpha_2 + \dots + \alpha_k = \eta$$

will obtain its optimal if and only if  $\alpha_1 = \alpha_2 = \dots = \alpha_k = \eta/k$ . This is desirable since we do not bias towards any predictors.

- (2) LAGS is a shrinkage estimator in the sense that

$$\sum_{i=1}^p \frac{|\hat{\beta}|_i}{|\beta_{OLS}|_i} \leq p$$

It is easily followed by the fact that  $l(\hat{\beta}; \lambda) \leq l(\beta_{OLS}; \lambda)$  and  $X^T(y - X\beta_{OLS}) = 0$ , thus

$$\frac{1}{n} \|X^T(y - X\hat{\beta})\|_1 + \lambda \sum_{i=1}^p \frac{|\hat{\beta}|_i}{|\beta_{OLS}|_i} \leq \lambda p$$

Suppose now that  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ , our next result is non-asymptotic, which claims that under the choice  $w_i = 1/|\beta_{OLS}|_i$ , LAGS can accurately identify the underlying model and estimate the coefficients.

**Theorem 4.2.** *Under the choice of  $w_i = 1/|\beta_{OLS}|_i$  and suppose that*

$$\min_{i \in \{1, 2, \dots, p_0\}} |\beta_i| > c\sqrt{2 \log p \sigma} \quad (18)$$

$$\frac{\|C_n\|_\infty}{\gamma_n} \leq M \quad (19)$$

where  $\gamma_n = \min_{s \in \Omega} \{ \|[C_{11}^n, C_{12}^n](s_1, s_2)^T\|_\infty \}$  as in (16). If  $\xi > 0$  satisfies  $(c - \xi)/\xi > M$ , then we can choose

$$\xi\sqrt{2 \log p \sigma} \|C_n\|_\infty \leq \lambda \leq (c - \xi)\sqrt{2 \log p \sigma} \gamma_n \quad (20)$$

such that

$$\hat{\beta}_i = 0, i = p_0 + 1, \dots, p \quad (21)$$

$$\hat{\beta}_i = (\beta_{OLS})_i, i = 1, \dots, p_0 \quad (22)$$

$$\|\hat{\beta} - \beta\|_2^2 \leq 2\xi^2 \cdot p_0 \cdot \log p \cdot \sigma^2 \quad (23)$$

are satisfied with probability at least  $(1 - \pi^{-1/2} \xi^{-1} (n \log p)^{-1/2} \kappa p^{-n\xi^2/\kappa^2})^p$ , where  $\kappa$  is a constant dependent on  $C_n^{-1/2}$ .

In words, the nonzero coefficients should significantly stand above the noise as indicated by (18) and  $\|C_n\|_\infty/\gamma_n$  is uniformly bounded by  $M$  as in (19),  $\xi$  is chosen to make the set (20) nonempty. If all these conditions are satisfied, LAGS identifies the correct variables and only these with large probability. Moreover, the coefficients of identified variables are set to be the OLS estimate, which is analogous to hard thresholding. The accuracy of LAGS is quantified by (23), the mean squared error is proportional to the true number of variables times the variance of the noise with the logarithmic factor— $\log p$ , which is unavoidable since we do not have the oracle to know the set of true predictors in advance [3, 2].

## 5 Proofs of Theorems

### 5.1 Proof of Theorem 3.1

*Proof.* LAGS is a convex program, thus it obtains its minimum at some  $\hat{\beta}$ . By the theory of convex optimization, the subgradient at  $\hat{\beta}$  is zero

$$\begin{aligned} \nabla l(\hat{\beta}) &= -\frac{X^T X}{n} \text{sign}(X^T y - X^T X \hat{\beta}) \\ &\quad + \lambda_n (w_1 \text{sign}(\hat{\beta}_1), \dots, w_{p_0} \text{sign}(\hat{\beta}_{p_0}), w_{p_0+1} \text{sign}(\hat{\beta}_{p_0+1}), \dots, w_p \text{sign}(\hat{\beta}_p))^T \\ &= 0 \end{aligned} \tag{24}$$

where  $\text{sign}(x) = \frac{x}{|x|}$  for  $x \neq 0$  and  $\text{sign}(x) \in [-1, 1]$  for  $x = 0$ . We take  $\tilde{b} = \|C\|_\infty$ . Since  $b = \lim \lambda_n b_n > \|C\|_\infty$ , for any  $\delta \in (0, (b - \tilde{b}))$ , there exists  $N_1(\delta)$  and  $n > N_1(\delta)$  such that  $\|C\|_\infty + \delta < \lambda_n b_n$ . Similarly, since  $C_n \rightarrow C$  as in condition (b), there exists  $N_2(\delta)$  and  $n > N_2(\delta)$  such that  $\|C_n\|_\infty \leq \|C\|_\infty + \delta$ .

The optimal condition (24) implies

$$\left[ \frac{X^T X}{n} \text{sign}(X^T y - X^T X \hat{\beta}) \right]_i = \lambda_n w_i \text{sign}(\hat{\beta}_i) \tag{25}$$

Notice that when  $p_0 + 1 \leq i \leq p$ ,

$$\lambda_n b_n \leq \lambda_n w_i$$

and

$$\left[ \frac{X^T X}{n} \text{sign}(X^T y - X^T X \hat{\beta}) \right]_i \leq \left\| \frac{X^T X}{n} \right\|_\infty$$

Hence

$$\lambda_n b_n |\text{sign}(\hat{\beta}_i)| \leq \left\| \frac{X^T X}{n} \right\|_\infty \tag{26}$$

for  $n > N(\delta) = \max\{N_1(\delta), N_2(\delta)\}$ , we have  $|\text{sign}(\hat{\beta}_i)| < 1$ , which implies  $\hat{\beta}_i = 0$ .

Analogously, we take  $\tilde{a} = \gamma$ , where  $\gamma$  is defined in Section 3; when  $1 \leq i \leq p_0$  and since  $a = \lim \lambda_n a_n < \tilde{a}$ , for any  $\delta' \in (0, (\tilde{a} - a))$ , there exists  $N(\delta')$

and  $n > N(\delta')$  such that  $\min_{s \in \Omega} \{ \| [C_{11}^n, C_{12}^n] s \|_\infty \} > \gamma - \delta'$  and  $\lambda_n a_n \leq \gamma - \delta'$ . Choose  $n > N = \max\{N(\delta), N(\delta')\}$ , we have  $\hat{\beta}^{(2)} = 0$ , then

$$\begin{aligned} \text{sign}(X^T y - X^T X \hat{\beta}) &= \text{sign} \left( \begin{pmatrix} X^{(1)T} y \\ X^{(2)T} y \end{pmatrix} - \begin{pmatrix} X^{(1)T} X^{(1)} \hat{\beta}^{(1)} \\ X^{(2)T} X^{(1)} \hat{\beta}^{(1)} \end{pmatrix} \right) \\ &= \text{sign} \left( \begin{pmatrix} X^{(1)T} \epsilon \\ X^{(2)T} \epsilon \end{pmatrix} - \begin{pmatrix} X^{(1)T} X^{(1)} (\hat{\beta}^{(1)} - \beta^{(1)}) \\ X^{(2)T} X^{(1)} (\hat{\beta}^{(1)} - \beta^{(1)}) \end{pmatrix} \right) \end{aligned} \quad (27)$$

We claim that  $X^{(1)T} \epsilon - X^{(1)T} X^{(1)} (\hat{\beta}^{(1)} - \beta^{(1)}) = 0$ .

If it is not true, then

$$s' = \text{sign}(X^T \epsilon - X^T X (\hat{\beta} - \beta)) \in \Omega$$

which upon combining with (25) gives

$$\gamma - \delta' < \| [C_{11}, C_{12}] s' \|_\infty < \lambda_n a_n \quad (28)$$

contradicts with our choice of  $n$ .

We impose the superscript  $n$  on  $\hat{\beta}$  to make it explicitly dependent on  $n$ . Therefore, for  $n > N$ ,

$$\frac{1}{n} \| X^{(1)T} \epsilon - X^{(1)T} X^{(1)} (\hat{\beta}^{n(1)} - \beta^{(1)}) \|_1 = 0$$

Combining  $\mathbb{E}(\epsilon) = 0$  and the law of large numbers,

$$\frac{X^{(1)T} \epsilon}{n} \rightarrow 0$$

by condition (b)

$$\frac{X^{(1)T} X^{(1)}}{n} \rightarrow C_{11}$$

hence,

$$\hat{\beta}^{n(1)} \rightarrow \beta^{(1)}$$

which implies LAGS is consistent.

The pseudo-hard thresholding property holds for  $n > N$  as well. In fact, as shown above, if inequalities (26) and (28) are satisfied, then we have

$$\hat{\beta}^{(1)} = (X^{(1)T} X^{(1)})^{-1} X^{(1)T} y \quad (29)$$

and

$$\hat{\beta}^{(2)} = 0 \quad (30)$$

Therefore, for any sequence  $\{\lambda_n w_1, \lambda_n w_2, \dots, \lambda_n w_p\}$ , if  $\lambda_n a_n \leq \tilde{a} - \delta'$  and  $\lambda_n b_n \geq \tilde{b} + \delta$ , then the LAGS solutions are the same.  $\square$

## 5.2 Proof of Theorem 4.1

*Proof.* The normal equation is

$$X^T X \beta = X^T y \quad (31)$$

where  $y = X^{(1)} \beta^{(1)} + \epsilon$ . By condition (a), (b),

$$\frac{1}{n} X^T X \rightarrow C, \frac{1}{n} X^T \epsilon \rightarrow 0, \frac{1}{n} X^T X^{(1)} \beta^{(1)} \rightarrow \begin{pmatrix} C_{11} \\ C_{21} \end{pmatrix} \beta^{(1)}$$

so

$$\hat{\beta}^{(1)} \rightarrow \beta^{(1)}, \hat{\beta}^{(2)} \rightarrow \beta^{(2)} = 0$$

Hence, when  $n$  is large enough, we can choose  $\lambda_n$ , such that  $\lambda_n / |\beta_{OLS}|_i < \gamma$  for  $i = 1, 2, \dots, p_0$  and  $\lambda_n / |\beta_{OLS}|_i > \|C\|_\infty$  for  $i = p_0 + 1, \dots, p$ .  $\square$

## 5.3 Proof of Theorem 4.2

In order to prove this theorem, we need the following lemmas.

**Lemma 5.1.** *If  $z \sim \mathcal{N}(0, \Sigma)$ ,  $z \in \mathcal{R}^p$  and  $\|\Sigma^{1/2}\|_\infty \leq c^{-1}$ , then the probability*

$$\mathbb{P}(\|z\|_\infty \leq t) \geq \left(1 - \frac{2\phi(ct)}{ct}\right)^p$$

where  $\phi(t) = (2\pi)^{-1/2} \exp(-t^2/2)$ .

*Proof.* Suppose  $v \sim \mathcal{N}(0, I)$  and  $z = \Sigma^{1/2} v$ ,

$$\int_{\|z\|_\infty \leq t} \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} z^T \Sigma^{-1} z\right) dz \quad (32)$$

$$= \int_{\|\Sigma^{1/2} v\|_\infty \leq t} \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} v^T v\right) dv \quad (33)$$

$$\geq \int_{\|v\|_\infty \leq ct} \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} v^T v\right) dv \quad (34)$$

$$\geq \left(1 - \frac{2\phi(ct)}{ct}\right)^p \quad (35)$$

Here we apply the fact that  $\int_t^\infty \phi(t)dt \leq \phi(t)/t$  for  $t > 0$ .  $\square$

**Lemma 5.2.** *Suppose  $\|C_n^{-1/2}\|_\infty \leq \kappa$ , then  $|\beta_{OLS}|_i \geq (c-\xi)\sqrt{2\log p}\sigma$  for  $i = 1, \dots, p_0$  and  $|\beta_{OLS}|_i \leq \xi\sqrt{2\log p}\sigma$  for  $i = p_0 + 1, \dots, p$  are satisfied with probability at least  $\left(1 - \frac{2\phi(\xi\sqrt{2n\log p}/\kappa)}{\xi\sqrt{2n\log p}/\kappa}\right)^p = (1 - \pi^{-1/2}\xi^{-1}(n\log p)^{-1/2}\kappa p^{-n\xi^2/\kappa^2})^p$ .*

*Proof.* Since  $y = X^{(1)}\beta^{(1)} + \epsilon$

$$\beta_{OLS} = (X^T X)^{-1} X^T y \quad (36)$$

$$= \begin{pmatrix} \beta^{(1)} \\ 0 \end{pmatrix} + \begin{pmatrix} X^{(1)T} X^{(1)} & X^{(1)T} X^{(2)} \\ X^{(2)T} X^{(1)} & X^{(2)T} X^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} X^{(1)T} \epsilon \\ X^{(2)T} \epsilon \end{pmatrix} \quad (37)$$

$$(38)$$

Denote  $\zeta = \begin{pmatrix} X^{(1)T} X^{(1)} & X^{(1)T} X^{(2)} \\ X^{(2)T} X^{(1)} & X^{(2)T} X^{(2)} \end{pmatrix}^{-1} \begin{pmatrix} X^{(1)T} \epsilon \\ X^{(2)T} \epsilon \end{pmatrix}$ , which is a Gaussian random vector with mean 0 and variance  $C_n^{-1}\sigma^2/n$ . Hence by applying Lemma 5.1 and noting  $\|C_n^{-1/2}\sigma/\sqrt{n}\|_\infty \leq \kappa\sigma/\sqrt{n}$

$$\mathbb{P}(\|\zeta\|_\infty \leq \xi\sqrt{2\log p}\sigma) \geq \left(1 - \frac{2\phi(\xi\sqrt{2n\log p}/\kappa)}{\xi\sqrt{2n\log p}/\kappa}\right)^p$$

$\square$

Theorem 4.2 is a consequence of the two lemmas.

*Proof.* Since  $X^T(y - X\beta_{OLS}) = 0$ , it implies

$$l(\beta) = \frac{1}{n}\|X^T y - X^T X\beta\|_1 + \lambda \sum_{i=1}^p w_i |\beta_i| = \frac{1}{n}\|X^T X(\beta_{OLS} - \beta)\| + \lambda \sum_{i=1}^p w_i |\beta_i|$$

Choosing  $\lambda$  satisfying (20) and applying Lemma 5.1 and Lemma 5.2, we take the subgradient of  $l(\beta)$  and use the same arguments as in the proof of Theorem 3.1. With large probability, we have

$$\hat{\beta}_i = 0, i = p_0 + 1, \dots, p$$

and

$$\left(X^{(1)T} X^{(1)}, X^{(1)T} X^{(2)}\right)(\hat{\beta}^{(1)} - \beta_{OLS}^{(1)}) = 0$$

which implies

$$\hat{\beta}_i = (\beta_{OLS})_i, i = 1, \dots, p_0$$

Hence

$$\|\hat{\beta} - \beta\|_2^2 = \left\| \begin{pmatrix} \beta_{OLS}^{(1)} \\ 0 \end{pmatrix} - \begin{pmatrix} \beta^{(1)} \\ 0 \end{pmatrix} \right\|_2^2 \quad (39)$$

$$= \left\| \begin{pmatrix} \zeta^{(1)} \\ 0 \end{pmatrix} \right\|_2^2 \quad (40)$$

$$\leq 2\xi^2 \cdot p_0 \cdot \log p \cdot \sigma^2 \quad (41)$$

by  $\|\zeta\|_\infty \leq \xi\sqrt{2\log p}\sigma$ . □

## 6 Computation and Implementation

LAGS is a convex program, thus global minimum is assured. There are a lot of algorithms available to solve (5), such as the subgradient method, interior-point methods. However, like Dantzig Selector [2], LAGS can also be reformulated as a linear program.

Denote  $|(X^T(y - X\beta))_i| = u_i$ ,  $i = 1, \dots, p$ ;  $|\beta_i| = v_i$ ,  $i = 1, \dots, p$ . Then solving LAGS is equivalent to solve the following linear program with inequality constraints

$$\min_{u, v, \beta} \sum_{i=1}^p u_i + \lambda \sum_{i=1}^p w_i v_i \quad (42)$$

subject to

$$-u \leq X^T(y - X\beta) \leq u \quad (43)$$

$$-v \leq \beta \leq v \quad (44)$$

This linear program has  $3p$  unknowns and  $4p$  constraints. When  $p$  is relatively small, e.g., less than 100, solving linear program is more efficient; when  $p$  gets large, solving the convex program directly is recommended.

If the computing environment contains the routine of solving regression under least absolute deviance criterion (LAD). LAGS can also be passed into the routine by treating an augmented samples and responses. Let us define  $(y^*, X^*)$ , where  $(y_i^*, x_i^*) = ((X^T y)_i, (X^T X)_i)$  for  $1 \leq i \leq p$ ;  $(y_{p+i}^*, x_{p+i}^*) =$



$(0, \lambda w_i \mathbf{e}_j)$  for  $1 \leq i \leq p$ ,  $\mathbf{e}_j$  is the  $j$ th row of identity matrix. Then it can be verified that

$$\hat{\beta} = \arg \min \|\mathbf{y}^* - \mathbf{X}^* \beta\|_1$$

So LAGS can be solved without much programming effort.

In most applications, we need to run a sequence of sparsity parameter  $\lambda$  and choose the optimal one based on some criteria such as cross-validation. An efficient way to accomplish this is to pass the previous solution as the “warm start” for the new value of  $\lambda$ . The justification of this technique is that in simplex algorithm, each iteration tries to find a better candidate solution at the vertices adjacent to the current one, if the difference between  $\lambda$ s are small, the new minimum should be close to the previous one, hence the new solution can be found in a few iterations.

## 7 Numerical Simulation and Example

### 7.1 Prostate Cancer Data

For the sake of illustrating the Pseudo-Hard Thresholding property, we study the simple yet popular example—Prostate Cancer data [15, 16]. The response—logarithm of prostate-specific antigen (lpsa) is regressed on log(cancer volume) (lcavol), log(prostate weight) (lweight), age, the logarithm of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion(svi), log(capsular penetration) (lcp), Gleason score (gleason) and percentage Gleason score 4 or 5(pgg45).

The data set is divided into two parts: a training set of 67 observations and a test set of 37 observations. It clearly shows in Figure 3 that the coefficients profiles of LAGS and  $l_0$  penalty are quite similar. As  $\lambda$  increases, the predictors are excluded from the model with the same order at almost the same  $\lambda$ s. The discrete selection processes of LAGS and  $l_0$  penalty are indicated by the jumps and constant segments in the profiles: the jumps means the predictors are either included in the model or not; while the constant segments means the coefficients of the included predictors are unchanged even though  $\lambda$  increases. Another interesting observation of the profiles is that the roles of some predictors are downplayed when there are many predictors included, which are mainly caused by the high correlation among them; however, as the included predictors are fewer, their roles are shown up and even increase—refer to the brown line in Figure 3, coefficient of lcavol. The

right panel is the prediction errors of LAGS and  $l_0$  penalty on the test set respectively, they are also piecewise constant functions of  $\lambda$ .

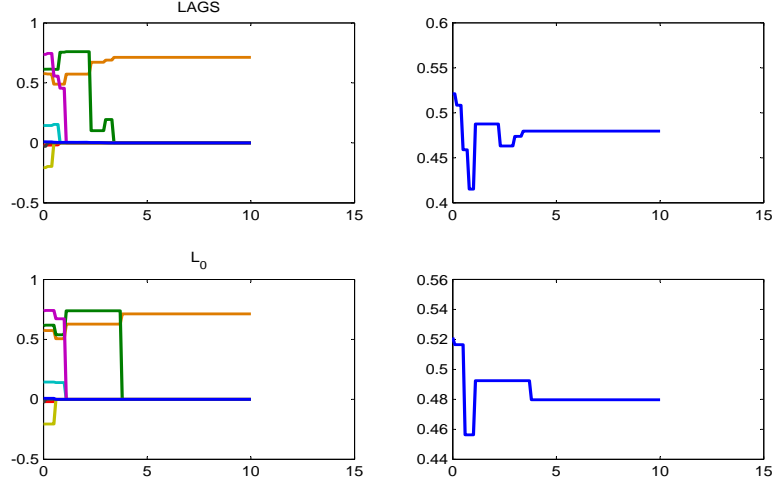


Figure 3: Coefficients profiles as a function of sparsity parameter  $\lambda$ . There are eight predictors, all the predictors are first centered and standardized before passing to the solvers, then the outputs of solvers are transformed back to the original scale. The right panel is the prediction error for the test data.

In contrast, the continuous shrinkage property of Lasso are demonstrated in Figure 4. The *general* trend of the coefficients is decreasing as  $\lambda$  increases (this is not true in general, there exist examples that some coefficients increase even though  $\lambda$  increases). The prediction error is also larger than that of LAGS and  $l_0$  penalty; but care should be taken that this argument can not be generalized; on the contrary, the discrete selection process usually exhibits higher variability hence higher prediction error.

## 7.2 Diabetes Data

The diabetes data are studied in [6, 5]. There are total 442 patients(samples) and 10 predictors in the model. We fit a linear model on this data set. Since LAGS tries to minimize the gradient directly by the  $l_1$ , it is insightful to

compare the gradients between LAGS and Lasso<sup>1</sup>. We have both computed the LAGS and Lasso solution with 5-fold cross validation as in Table 1; since LAGS sometimes demonstrates a little higher variability, we pick the most parsimonious model within 45%-50% standard error of the minimum instead of the “one-standard error” rule. LAGS has 4 nonzero coefficients while Lasso has 5. Moreover, we see in the table that the absolute inner product of predictors with the residue of LAGS is very sparse by its effort to minimizing the gradient directly, whereas that of Lasso is much denser and satisfies

$$X_j^T(y - X\beta^{lasso}) = \lambda \cdot \text{sign}(\beta_j^{lasso}), \quad \forall \beta_j^{lasso} \neq 0$$

and

$$|X_j^T(y - X\beta^{lasso})| \leq \lambda, \quad \forall \beta_j^{lasso} = 0$$

It is also notable that the magnitudes of the nonzero coefficients of LAGS are larger than that of Lasso. The reason is that Lasso both shrinks and selects, it tries to relax the penalty on relevant coefficients, as a consequence, some important predictors are downplayed by sharing their weights to others and selected model is relatively dense. This argument can be further verified by comparing the  $\|\beta\|_1$ :  $\|\beta^{lasso}\|_1 \approx 1335$  and  $\|\beta^{lags}\|_1 \approx 1617$ . Table 2 summarizes the correlations between the predictors and residue, which shows that LAGS enables some predictors to be exactly orthogonal to the residue. This simple data set shows the superiority of LAGS over Lasso.

---

<sup>1</sup>Lasso solution is computed by R package `glmnet` [11]

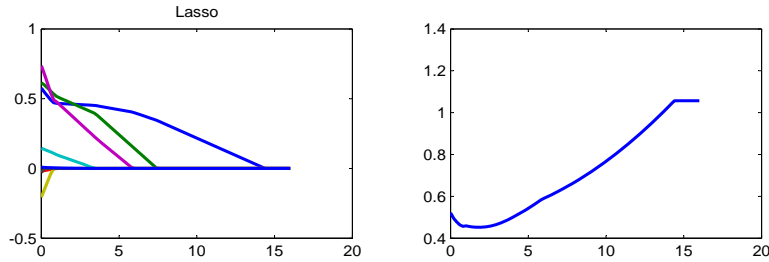


Figure 4: Coefficients profiles as a function of sparsity parameter  $\lambda$  by the Lasso.

Variable $j$	LAGS		Lasso	
	$X_j^T(y - X\beta)$	$\hat{\beta}_j$	$X_j^T(y - X\beta)$	$\hat{\beta}_j$
1	-27.9927	0.0000	14.5431	0.0000
2	-134.9231	0.0000	-111.7310	-33.3383
3	0.0000	604.7797	111.7310	508.1903
4	0.0000	268.1098	111.7310	210.3536
5	-53.2906	-133.8965	-55.5267	0.0000
6	0.0000	0.0000	-54.0219	0.0000
7	119.4116	0.0000	111.7310	138.8478
8	73.3477	0.0000	66.2507	0.0000
9	0.0000	609.8394	111.7310	444.5615
10	39.7171	0.0000	101.9332	0.0000
Mean Squared Error	3021		3044	

Table 1: The gradient  $X_j^T(y - X\beta)$  and coefficient  $\beta_j$  in each coordinate on the diabetes data with 10 predictors by LAGS and Lasso respectively.

	correlations between the predictors and residue									
LAGS	-0.02	-0.12	0.00	0.00	-0.05	0.00	0.10	0.06	0.00	0.03
Lasso	0.01	-0.10	0.10	0.10	-0.05	-0.05	0.10	0.06	0.10	0.09

Table 2: The correlations with the residue:  $X_j \cdot \text{res} / \|X_j\|_2 \|\text{res}\|_2$

### 7.3 Simulated Data

We compare the simulation performance of LAGS and Sparsenet<sup>2</sup> [13] and Lasso [16] with regard to training error, prediction error and number of non-zero coefficients in the model. We assume that the predictors and errors are Gaussian distributed. If  $X \sim \mathcal{N}(0, \Sigma)$  and  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ , then the Signal-to-Noise Ratio (SNR) is defined as

$$\text{SNR} = \frac{\sqrt{\beta^T \Sigma \beta}}{\sigma}$$

We take  $\Sigma = \Sigma(\rho) \in \mathbf{R}^{p \times p}$  with 1's on the diagonal and  $\rho$  on the off-diagonal. We generate two data sets with  $\text{SNR} = 2$ ,  $\rho = 0.2$  and  $\text{SNR} = 3$ ,  $\rho = 0.4$  respectively, the sample sizes  $n$  are both 2000 and  $\beta = (30, 29, 28, \dots, 1, 0_{970})$ ,  $p = 1000$ . In order to evaluate the performances of these three algorithms when  $p \gg n$ , we split the two data sets into training set with 500 samples and testing set with 1500 samples. Before passing the training set into the three algorithms, we first standardize the predictors, then choose the sparsity parameter  $\lambda$  by 10-fold cross-validation; since  $p \gg n$ , the weights for LAGS are set to be the inverse of ridge estimate with  $\phi = 0.2$  (since each predictor is standardized, the diagonals of  $X^T X$  are 1s,  $\phi$  is chosen quite arbitrarily from  $(0, 1)$ ). The results are summarized in Table 3, we can observe that Lasso solution is overly dense, whereas Sparsenet solution is overly sparse; LAGS stays between the two and is closer to the true model. In Table 4, the data sets are split into training set with 1500 samples and testing set with 500 samples, the weights are inverse of OLS estimates,  $\lambda$  is again chosen by 10-fold cross-validation. In both simulations, the prediction errors of LAGS are slightly larger than that of Sparsenet because of the variability caused by discreteness; however, LAGS tends to discover the true models while Sparsenet tends to discover the sparser ones.

## 8 Discussion

### 8.1 Adapted version of Dantzig Selector

In section 4, we argue that one possible explanation of DS in [6] is the uniformity of the weight put on each estimator. We adopt the choice of

---

<sup>2</sup>The data sets and Sparsenet solution are computed by R package `sparsenet` [13]

	LAGS	Sparsenet	Lasso
$\rho = 0.2$ SNR = 2	# nonzeros: 33 Training Error: 1915.4 Testing Error: 2258.3	# nonzeros: 20 Training Error: 1874.3 Testing Error: 2228.3	# nonzeros: 92 Training Error: 1624.3 Testing Error: 2505.5
$\rho = 0.4$ SNR = 3	# nonzeros: 30 Training Error: 710.3 Testing Error: 785.5	# nonzeros: 22 Training Error: 714.2 Testing Error: 772.0	# nonzeros: 112 Training Error: 584.1 Testing Error: 938.8

Table 3:  $n = 500$ ,  $p = 1000$ , # test set = 1500. The number of true predictors is 30.

	LAGS	Sparsenet	Lasso
$\rho = 0.2$ SNR = 2	# nonzeros: 29 Training Error: 1900.7 Testing Error: 2164.9	# nonzeros: 26 Training Error: 1854.1 Testing Error: 2064.4	# nonzeros: 97 Training Error: 1811.9 Testing Error: 2274.6
$\rho = 0.4$ SNR = 3	# nonzeros: 29 Training Error: 654.3 Testing Error: 683.2	# nonzeros: 26 Training Error: 594.1 Testing Error: 682.4	# nonzeros: 122 Training Error: 607.5 Testing Error: 760.7

Table 4:  $n = 1500$ ,  $p = 1000$ , # test set = 500. The number of true predictors is 30.

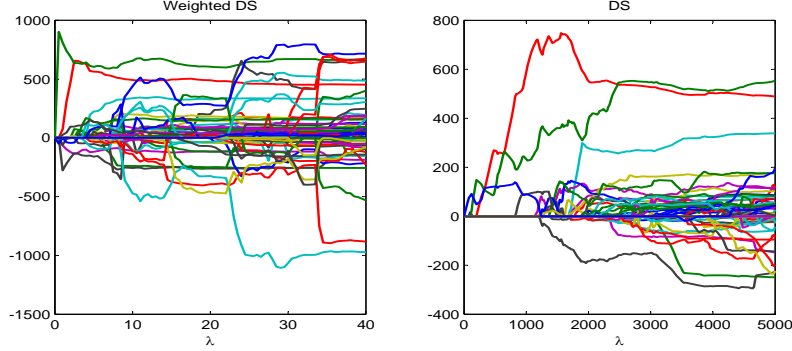


Figure 5: The coefficient profiles of the weighted-DS and DS on diabetes data with interaction terms.

weights as with LAGS. Thus, we need to solve the weighted version of DS

$$\hat{\beta} = \arg \min \|X^T(y - X\beta)\|_{\infty} \quad \text{subject to} \quad \sum_{i=1}^p \frac{|\beta_i|}{|\beta_{OLS}|_i} \leq t$$

We test our idea on the extended diabetes data where the interactions are included. The left panel of Figure 5 is the weighted version, while the right panel is the original version which is also appeared in [6]. The erratic behavior of DS is relatively mitigated with the weighted DS.

## 8.2 Contributions in this paper

In this paper, we propose a new algorithm for linear regression—LAGS and introduce the “pseudo-hard thresholding” properties which mimics the  $l_0$  regularization. Under mild conditions, we have proved that asymptotically, LAGS is consistent for model selection and parameter estimation. Since the strength of the variable selection algorithms lies in its finite sample performance, we have also established the nonasymptotic theorem which shows that with large probability, LAGS can discover the true model and the error of the estimated parameters is controlled under the noise level. In the proofs of these theorems, we emphasize that the weights on the parameters play a critical role for the effectiveness of LAGS.

LAGS is a re-weighted regularization method which is first discussed in adaptive Lasso [19], it can be also interpreted as a multistage procedure:

first, we provide a very coarse estimate of the parameters of interest; second, based on this estimate, we are able to seek much better ones. Letting the regularization part depend on the data set makes the theories much more difficult; however, it usually results in better performance.

The subject of variable selection in linear models has large bodies of literature. The efforts are mainly divided into two streams: on the one hand, the discrete selection procedures of  $l_0$  penalty methods such as AIC, BIC are shown to enjoy many nice properties, but they are highly impractical; on the other hand, the continuous shrinkage algorithms such as Lasso and LARS are computationally favorable but they do not have hard-thresholding property and the bias introduced by them is significant sometimes. Our work bridges the gap by pioneering the discrete selection process in the convex world. The attractive properties of LAGS indicate that in some applications, LAGS can be served as a surrogate for  $l_0$  penalty and an improved version of the continuous shrinkage methods.

There are a group of algorithms on the shelf to solve LAGS. However, since LAGS needs to compute a sequence of solutions like Lasso and Sparsenet, one possible future work is to develop efficient path algorithms such as glmnet and sparsenet.

## References

- [1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.
- [2] E. Candes and T. Tao. The dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics*, 35(6):2313–2351, 2007.
- [3] E.J. Candes and Y. Plan. Near-ideal model selection by  $l_1$  minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.
- [4] D.L. Donoho. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995.
- [5] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.



- [6] B. Efron, T. Hastie, and R. Tibshirani. Discussion: The dantzig selector: Statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics*, 35(6):2358–2364, 2007.
- [7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
- [8] J. Fan and J. Lv. Properties of non-concave penalized likelihood with  $np$ -dimensionality. *Manuscript*, 2008.
- [9] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.
- [10] D.P. Foster and E.I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, pages 1947–1975, 1994.
- [11] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.
- [12] C.L. Mallows. Some comments on  $c_p$ . *Technometrics*, 42(1):87–94, 2000.
- [13] R. Mazumder, J.H. Friedman, and T. Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.
- [14] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.
- [15] T.A. Stamey, JN Kabalin, JE McNeal, IM Johnstone, F. Freiha, EA Redwine, and N. Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of urology*, 141(5):1076, 1989.
- [16] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [17] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

- [18] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(2):2541, 2007.
- [19] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.
- [20] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.